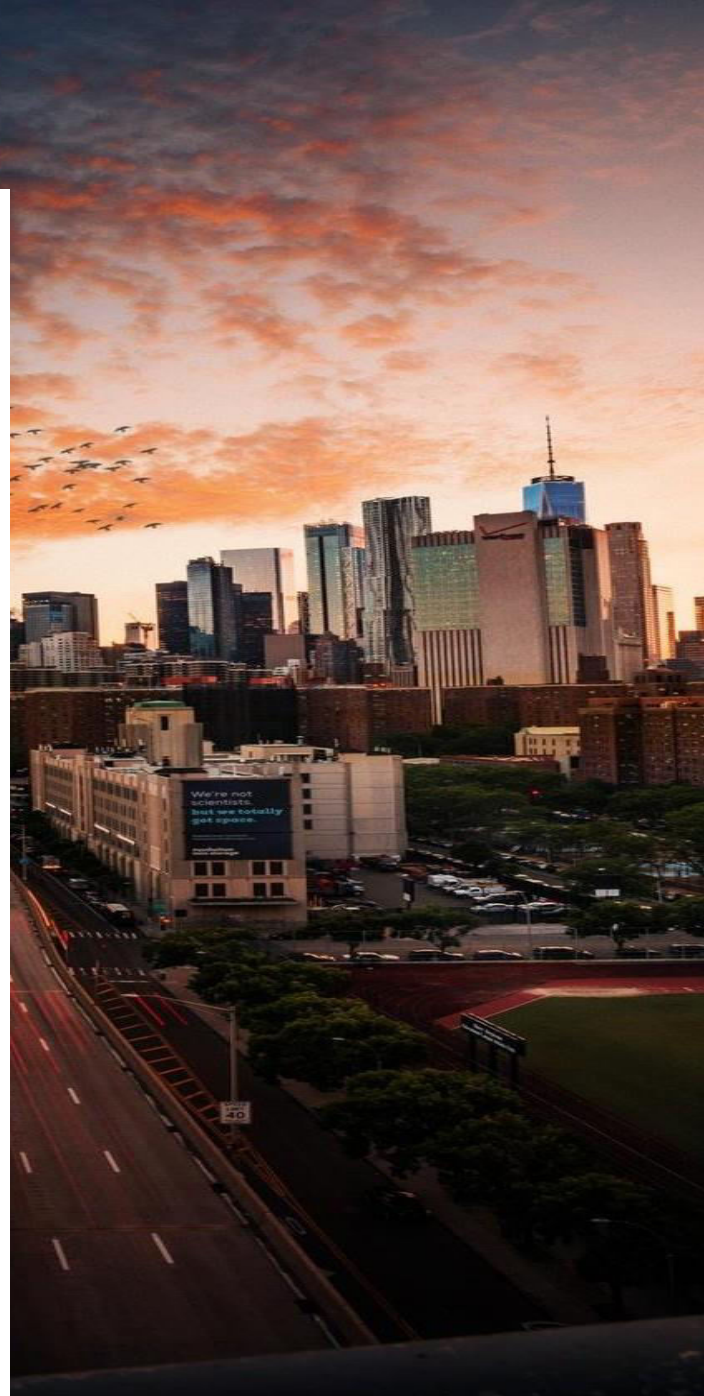# Coursera Capstone 2019

October 2019

Opening a New Restaurant in Mumbai, India

Authored by: Reshma Yemulwar

Opening a New Restaurant in Mumbai, India

Introduction

For many shoppers, visiting restaurants is a great way to relax and enjoy themselves during weekends and holidays. They can do grocery shopping, dine at restaurants, shop at the various fashion outlets, watch movies and perform many more activities. restaurants are like a one-stop destination for all types of food lovers. For retailers, the central location and the large crowd at the restaurants provides a great distribution channel to market their products and services. Property developers are also taking advantage of this trend to build more restaurants to cater to the demand. As a result, there are many restaurants in Mumbai and many more are being built. Opening restaurants allows property developers to earn consistent rental income. Of course, as with any business decision, opening a new restaurant requires serious consideration and is a lot more complicated than it seems. Particularly, the location of the restaurant is one of the most important decisions that will determine whether the restaurant will be a success or a failure.

Business Problem

The objective of this capstone project is to analyze and select the best locations in Mumbai, India in order to open a new restaurant. Using data science methodology and machine learning techniques like clustering, this project aims to provide solutions to answer the business question: In Mumbai, India, if a property developer is looking to open a new restaurant, where would you recommend that they open it?

Target Audience for this Project

This project is particularly useful to property developers and investors looking to open or invest in new restaurants in the financial capital city of India i.e. Mumbai.

Data

To solve the problem, we will need the following data:

- List of neighborhoods in Mumbai. This defines the scope of this project which is confined to the city of Mumbai, the financial capital city of India in Asia.
- Latitude and longitude coordinates of those neighborhoods. This is required in order to plot the map and also to get the venue data.
- Venue data, particularly data related to restaurants. We will use this data to perform clustering on the neighborhoods.

Sources of data and methods to extract them:

This Wikipedia page (https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Mumbai) contains a list of neighborhoods in Mumbai, with a total of 37 neighborhoods. We will use web scraping techniques to extract the data from the Wikipedia page, with the help of Python requests and beautifulsoup packages. Then we will get the geographical coordinates of the neighborhoods using Python Geocoder package which will give us the latitude and longitude coordinates of the neighborhoods.

After that, we will use Foursquare API to get the venue data for those neighborhoods. Foursquare has one of the largest database of 105+ million places and is used by over 125,000 developers. Foursquare API will provide many categories of the venue data, we are particularly interested in the restaurant category in order to help us to solve the business problem put forward. This is a project that will make use of many data science skills, from web scraping (Wikipedia), working with API (Foursquare), data cleaning, data wrangling, to machine learning (K-means clustering) and map visualization (Folium). In the next section, we will present the Methodology section where we will discuss the

steps taken in this project, the data analysis that we did and the machine learning technique that was used.

Methodology

Firstly, we need to get the list of neighborhoods in Mumbai. Fortunately, the list is available on the Wikipedia page (https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Mumbai). We have to do web scraping using Python requests and beautifulsoup packages to extract the list of neighborhoods data. However, this is just a list of names. We need to get the geographical coordinates in the form of latitude and longitude in order to be able to use Foursquare API. To do so, we will use the wonderful Geocoder package that will allow us to convert address into geographical coordinates in the form of latitude and longitude. After gathering the data, we will populate the data into a pandas DataFrame and then visualize the neighborhoods in a map using Folium package. This allows us to perform a sanity check to make sure that the geographical coordinate's data returned by Geocoder are correctly plotted in the city of Mumbai.

Next, we make use of Foursquare API to get the top 100 venues that are within a radius of 2000 meters. We need to register a Foursquare Developer Account in order to obtain the Foursquare ID and Foursquare secret key. We then make API calls to Foursquare passing in the geographical coordinates of the neighborhoods in a Python loop. Foursquare will return the venue data in JSON format and we will extract the venue name, venue category, venue latitude and longitude. With the data, we can check how many venues were returned for each neighborhood and examine how many unique categories can be curated from all the returned venues. Then, we will analyze each neighborhood by grouping the rows by neighborhood and taking the mean of the frequency of occurrence of each venue category. By doing so, we are also preparing the data for use in clustering. Since we are analyzing the "restaurant" data, we will filter the "restaurant" as venue category for the neighborhoods.

Finally, we will perform clustering on the data by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data

point to the nearest cluster, while keeping the centroids as restaurant as possible. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project. We will cluster the neighborhoods into 3 clusters based on their frequency of occurrence for "restaurant". The results will allow us to identify which neighborhoods have higher concentration of restaurants while which neighborhoods have fewer number of restaurants. Based on the occurrence of restaurants in different neighborhoods, it will help us to answer the question as to which neighborhoods are most suitable to open new restaurants.

Results

The results from the k-means clustering show that we can categorize the neighborhoods into 3 clusters based on the frequency of occurrence for "restaurant":

- Cluster 0: Neighborhoods with low number to no existence of restaurants
- Cluster 1: Neighborhoods with high concentration of restaurants
- Cluster 2: Neighborhoods with moderate number of restaurants

The results of the clustering are visualized in the map below with cluster 0 in red color, cluster 1 in purple color, and cluster 2 in mint green color.
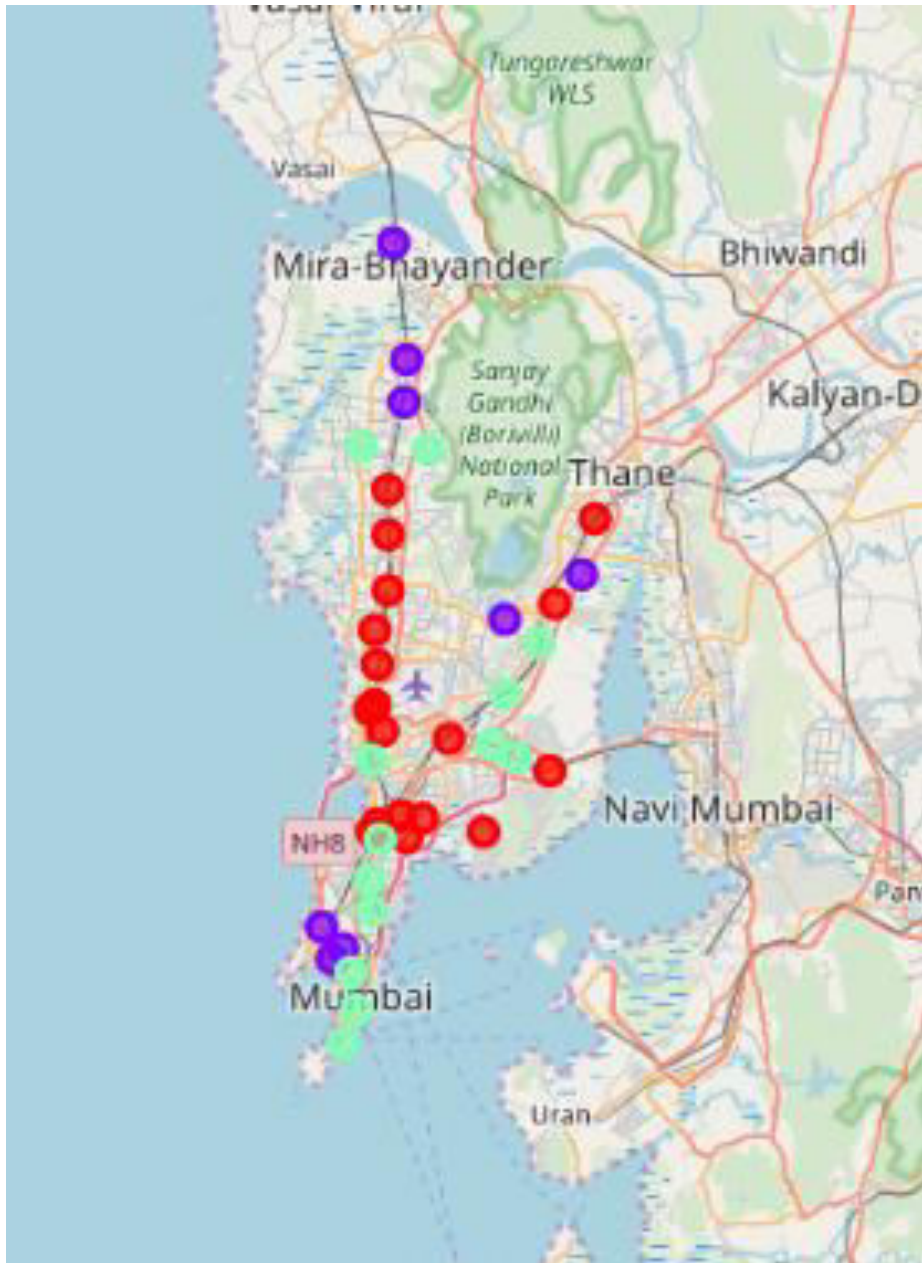
Discussion

Most of the restaurants are concentrated in the Northern arears of Mumbai city, with the highest number in cluster 1 and moderate number in cluster 2. On the other hand, cluster 0 has very low number of restaurant in the neighborhoods. This represents a great opportunity and high potential areas to open new restaurants as there is very

little to no competition from existing restaurants. Meanwhile, restaurants in cluster 1 are likely suffering from intense competition due to oversupply and high concentration of restaurants. From another perspective, this also shows that the oversupply of restaurants mostly happened in the developed parts like Thane in Mumbai city, with the suburb areas like South Mumbai still have very few restaurants. Therefore, this project recommends property developers to capitalize on these findings to open new restaurants in neighborhoods in cluster 0 with little to no competition. Property developers with unique selling propositions to stand out from the competition can also open new restaurants in neighborhoods in cluster 2 with moderate competition. Lastly, property developers are advised to avoid neighborhoods in cluster 1 which already have high concentration of restaurants and suffering from intense competition.

Limitations and Suggestions for Future Research

In this project, we only consider one factor i.e. frequency of occurrence of restaurants, there are other factors such as population and income of residents that could influence the location decision of a new restaurant. However, to the best knowledge of mine, such data are not available to the neighborhood level required by this project. Future research could devise a methodology to estimate such data to be used in the clustering algorithm to determine the preferred locations to open a new restaurant. In addition, this project made use of the free Sandbox Tier Account of Foursquare API that came with limitations as to the number of API calls and results returned. Future research could make use of paid account to bypass these limitations and obtain more results.

Conclusion

In this project, we have gone through the process of identifying the business problem, specifying the data required, extracting and preparing the data, performing machine learning by clustering the data into 3 clusters based on their similarities, and lastly providing recommendations to the relevant stakeholders i.e. property developers and investors regarding the best locations to open a new restaurant. To answer the business question that was raised in the introduction section, the answer proposed by

this project is: The neighborhoods in cluster 0 are the most preferred locations to open a new restaurant. The findings of this project will help the relevant stakeholders to capitalize on the opportunities on high potential locations while avoiding overcrowded areas in their decisions to open a new restaurant.